СЕКЦИЯ № 3
ГЕОИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Председатель секции:
Марков Николай Григорьевич, д. техн. н., профессор, зав. каф. ВТ ИК ТПУ.
Секретарь секции:
Кудинов Антон Викторович, канд. техн. н., доцент каф. ВТ ИК ТПУ.

УДК 004

ОБРАБОТКА РЕШЕНИЙ СУДОВ ТОМСКОЙ ОБЛАСТИ И ГОРОДА ТОМСКА С ПОМОЩЬЮ ТЕХНОЛОГИЙ OLAP И DATA MINING

Хлопонин А.А., Паршина Д.М.
Научный руководитель: Кудинов А.В.

Национальный Исследовательский Томский политехнический университет,
634050, Россия, г. Томск, пр. Ленина, 30
E-mail: alex@diplux.com, sirena13@sibmail.com

The article is intended to analyze various data obtained from websites of regional and district Tomsk courts via advanced analytic technologies such as OLAP and Data Mining. The process of comparing structure open documents and their parsing using Python and NoSQL databases are considered in details. Near-duplicates and shingling, as well as regular expressions stand for analyzing and comparing texts, sentences and words. Due to these algorithms, the issue relating to extraction of necessary units can be sorted out effectively and quite accurately.

Key words: *the Law field, Data mining, OLAP, Microsoft SQL Server Analysis Service, Elasticsearch, Kibana, HTML parser, Pyton, regular expressions, shingling, text analysis, relational database, NoSQL database.*

Ключевые слова: *судопроизводство, Data Mining, OLAP, Microsoft SQL Server Analysis Service, Elasticsearch, Kibana, парсинг HTML-страниц, Pyton, регулярные выражения, алгоритм шинглов, анализатор текста, реляционная база данных, NoSQL база данных.*

Введение

В наше время происходит интенсивное накопление огромных объёмов данных разного типа в различных предметных областях измеряемые в петабайтах, это в свою очередь даёт возможность решать задачи получения новых фактов, зависимостей и скрытых корреляций, а также позволяет решать некоторые аналитические задачи, такие как прогнозирование, проверка статистических гипотез, расчёт агрегатных показателей и т. д. В данной статье рассматриваются некоторые технологии и алгоритмы для извлечения и анализа данных на примере судопроизводства в г. Томске и Томской области.

Постановка задачи

Основными задачами являются извлечение данных из открытых источников, официально предоставляемых судами РФ (официальные сайты судов РФ, сайты организаций, занимающихся обработкой делопроизводства судов РФ) и их анализ с помощью технологий OLAP, Data Mining и Text Mining.

Поставленную задачу можно условно разделить на ряд следующих подзадач:

- Анализ источников данных судебных дел Томских судов.
- Создание информационной модели: выделение основных параметров судебных дел.
- Анализ источников данных и реализация анализатора текста для получения наборов данных судопроизводства.

Решение различных аналитических задач на полученном наборе данных с помощью технологий OLAP, Data Mining и Text Mining.

Анализ предметной области

Архив судебных актов Томских районных и областных судов состоит из административных, гражданских и уголовных дел. В зависимости от типа судебного решения можно выделить основные информационные объекты: постановления, решения, определения и приговоры. Каждый объект имеет общие атрибуты: номер дела, город, ФИО судьи, дата составления документа, название суда, нормативный акт (статья, часть, название), дата вступления в силу, тип наказания, тяжесть наказания, ФИО подсудимого, Название организации, ФИО адвоката, типы юридических лиц, фигурирующих в делах. В результате анализа предметной области была сформирована модель данных судопроизводства [1].

Извлечение данных

Инструментами для получения данных из решений судов являются терминал в операционной системе Linux и приложение, реализованное на языке Python3.4 с сопутствующими технологиями. Основным способом поиска и извлечения данных являются регулярные выражения, а также программные фильтры отсекающие некорректные данные пропущенные регулярными выражениями. Таким образом происходит формирование дополнительных векторов данных для каждого судебного решения, что в будущем позволит расширять применение алгоритмов майнинга данных. Все сформированные объекты импортируются в NoSQL базу данных Elasticsearch для последующих хранения и обработки. Выбор данного способа хранения информации связан с существенной гибкостью данного типа базы данных, возможностью дополнения и изменения данных и одновременного выполнения срезов данных.

Анализ данных

Технологии OLAP позволяет решать аналитические задачи, такие как расчёт статистических данных, а Data Mining рассматривает задачи интеллектуального анализа, статистической проверки гипотез, прогнозирование. К задачам получения агрегатных и статистических данных в судопроизводстве можно отнести следующие: подсчёт количества дел по видам; разделение решений по полу судей; подсчёт количества судебных актов, вступившие в силу в определённый период; подсчёт количества районных, областных судов, по превалированию уголовных, гражданских или административных нарушений; подсчёт процентного соотношения ведения дел по видам кодексов в районных и областных судах; выявление статистики по экономическим делам; выявление дел связанных с крупными фирмами и банками

на предмет возможных нарушений. Задачи такого типа можно эффективно решать при помощи технологии OLAP. К задачам интеллектуального анализа относятся: влияние пола судьи на решения по судебным делам; обнаружение явно выделяющихся дел из общих признаков, например, скорость принятия решений судьёй, несопоставимость приговора и тяжести преступления и т. д. К задачам прогнозирования можно отнести предсказание результата приговора по уголовному делу, учитывая следующие данные: степень тяжести совершенного уголовного преступления, предыдущая судимость обвиняемого, личность судьи, личность адвоката. Эти задачи можно решать при помощи технологии Data Mining. Для решения поставленных задач используется сервер хранения данных Elasticsearch и база данных Lucene лежащая в его основе, для графического отображения зависимостей и статистических показателей используется инструмент Kibana. Этот сервер предназначен для создания OLAP-кубов на основе нереляционных хранилищ данных. Построенный OLAP-куб содержит все данные необходимые для интеллектуального анализа [2].

Текущие результаты и перспективы

Был реализован парсер для извлечения содержимого из источников данных, анализатор текста, основанный на регулярных выражениях, а также была построена информационная модель данных судопроизводства. В дальнейшем планируется решение задач интеллектуального анализа с помощью технологий OLAP, Data Mining и Text Mining.

Список литературы

1. Щукова К.Б., Хлопонин А.А., Паршина Д.М. Извлечение и анализ данных о судопроизводстве в г. Томске с помощью технологий OLAP и DATA MINING // Технологии Microsoft в теории и практике программирования: сборник трудов XII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 25–26 Марта 2015. – Томск: Изд-во ТПУ, 2015. – С. 105–106.
2. Барсегян А.А., Куприянов М.С. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.

УДК 004

ИССЛЕДОВАНИЕ ИНВАРИАНТНЫХ КОРРЕЛЯЦИОННЫХ ФИЛЬТРОВ ДЛЯ ЗАДАЧ РАСПОЗНАНИЯ ОБРАЗОВ

Хлопонин И.А.
Научный руководитель: Болотова Ю.А.

Национальный Исследовательский Томский политехнический университет,
634050, Россия, г. Томск, пр. Ленина, 30
E-mail: khlopilia@gmail.com

To date, the correlation image recognition the most promising and widely used method of search, localization, identification of objects of complex shapes. The range of applications of such filters is very wide: the automatic diagnostics in medicine, biometric access systems, navigation systems, recognition, classification, and others.

Key words: *correlation filters, ASEF, MACE, Python, VanderLugt Filter, Minimum Average Correlation Energy Filter, Average of Synthetic Exact Filters, MNIST.*